

TITLE

SYSTEM AND METHOD FOR MULTI-LINGUAL SPEECH

RECOGNITION

BACKGROUND OF THE INVENTION

5 Field of the Invention

The present invention relates to speech recognition technology and in particular to a system and method for recognizing multiple languages in a single speech signal.

10 Description of the Related Art

Currently, the main methods of recognition of a multi-lingual speech signal are described as follows. A recognition system constructed by several independent uni-lingual speech recognition subsystems
15 must select a language desired by users or computers in advance and designate a uni-lingual speech recognition subsystem to recognize speech signals. Obviously, the mentioned method only can deal with one language at one time, being unable to handle various
20 languages simultaneously. Strictly speaking, although the mentioned method includes different speech recognition subsystems, it does not provide multi-lingual speech recognition functionality.

A second method utilizes one language to simulate
25 other languages. That is, the phonetic transcriptions of one main language are utilized to simulate the pronunciation of other languages. For example, if Chinese is selected as the main language, then

phonetic transcriptions of Chinese will be used to simulate other languages, such as English or Japanese. As an example, "DVD" in English might be simulated by "dil bil dil" in Chinese. The second method can
5 partially resolve multi-lingual speech recognition problems. However, one difficulty of the second method is that many parts of speech cannot be simulated. Thus, an incomplete simulation may affect the whole recognition result. To give an example, the
10 "V" in English cannot be simulated properly by Chinese phonetic transcriptions, whereby improper simulation will affect the whole recognition result.

The third method utilizes global phonemes to label the speech of all languages and then refers to a
15 decision tree to classify and recognize the labeled speech. The third method can avoid the mentioned incomplete simulation problem, however, if there is a large amount of vocabulary, interference among different languages will be significant, degrading the
20 recognition result.

SUMMARY OF THE INVENTION

Accordingly, an object of the invention is to utilize diphone models to recognize a mixed multi-lingual speech signal.

25 The inventive method adopts cross-lingual diphone models to recognize the parts of the speech signal containing multiple languages and uni-lingual diphone models to recognize parts of containing only one. That is, only the parts transitioning between

languages will be recognized by cross-lingual diphone models, avoiding the interference of different languages.

A complete speech recognition system must be
5 trained by a large amount of speech data. Another
object of the invention is to provide an initial
integration method, applied in the initial
establishment of a speech recognition system. The
initial integration method integrates several
10 different independent trained diphone speech
recognition systems into one multi-lingual speech
recognition system, resolving initial establishment
problems of the speech recognition system.

To achieve the foregoing objects, the present
15 invention provides a system for multi-lingual speech
recognition. The inventive system includes a speech
modeling engine, a speech search engine, and a
decision reaction engine. The speech modeling engine
receives and transfers a mixed multi-lingual speech
20 signal into speech features. The speech search engine
locates and compares candidate data sets to find match
probability for candidate speech models. The decision
reaction engine selects the candidate speech commands
according to the match probability and generates a
25 speech command.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention can be more fully
understood by reading the subsequent detailed

description and examples with references made to the accompanying drawings, wherein:

Fig. 1 is a diagram of the system for multi-lingual speech recognition according to the present invention.

Fig. 2 is a diagram of establishment of the multi-lingual context-speech mapping data according to the present invention.

Fig. 3 is a diagram of establishment of the multi-lingual anti-models according to the present invention.

Fig. 4 is a detailed diagram of establishment of the multi-lingual anti-models according to the present invention.

Fig. 5 is a diagram illustrating cross-lingual data of the present invention according to one embodiment.

Fig. 6 is a diagram of an applied example of the present invention according to one embodiment.

Fig. 7 is a flowchart of the method for multi-lingual speech recognition according to the present invention.

DETAILED DESCRIPTION OF THE INVENTION

As summarized above, the present invention provides a system for multi-lingual speech recognition, including a speech modeling engine, a speech search engine, and a decision reaction engine.

The speech modeling engine receives a mixed multi-lingual speech signal, converts the multi-lingual speech signal into speech features.

The speech search engine receives the speech
5 features, locating and comparing candidate data sets referring to a multi-lingual model database. Each of the candidate data sets corresponds to the speech features and has several candidate speech models with match probability. The speech models are
10 characterized by diphone models. The speech search engine may refer to connecting sequences of the speech features and a speech rule database. The connecting sequences may follow some specific connection rules in particular application, such as ID or address.

15 The multi-lingual model database includes multi-lingual context-speech mapping data and multi-lingual anti-models.

The provided system also includes a multi-lingual baseform generation engine and a cross-lingual diphone
20 model generation engine to generate the multi-lingual context-speech mapping data. The multi-lingual baseform mapping engine compares multi-lingual query commands to obtain multi-lingual baseforms. The cross-lingual diphone model generation engine selects
25 and combines the multi-lingual baseforms into the multi-lingual context-speech mapping data.

The disclosed system further includes certain uni-lingual anti-model engines and an anti-model combination engine to generate the multi-lingual anti-
30 models. The uni-lingual anti-model generation engines

receive multi-lingual query commands to normalized and generate all uni-lingual anti-models for all needed languages. The anti-model combination engine combines the uni-lingual anti-models to generate the multi-
5 lingual anti-models.

The decision reaction engine, coupled to the speech search engine, selects resulting speech models corresponding to the speech features from the candidate speech models according to the match
10 probability and generates a speech command. Then, the decision reaction engine can produce reactions according to the recognized speech command.

Furthermore, the invention discloses a method for multi-lingual speech recognition.

15 First, the method transfers a mixed multi-lingual speech signal into speech features.

Next, the method locates and compares candidate data sets corresponding to the speech features by referring to a multi-lingual model database. Each of
20 the candidate data sets has candidate speech diphone models with corresponding match probability. Locating and comparison may be accomplished by referring to other rules or databases, such as the connecting sequences of the speech models or a speech rule
25 database.

The multi-lingual model database includes multi-lingual context-speech mapping data and multi-lingual anti-models.

The multi-lingual context-speech mapping data is
30 established by following generation steps. First, the

multi-lingual query commands are compared to obtain multi-lingual baseforms. The multi-lingual baseforms are then selected and combined into the multi-lingual context-speech mapping data. For instance, the
5 mentioned generation steps can execute some detailed changes according to pronunciation. Nevertheless, the detailed changes can be ignored in simpler recognition systems.

The above selection and combination is
10 accomplished by certain steps. First, the left contexts of the multi-lingual baseforms are fixed, and the right contexts of the multi-lingual baseforms are mapped to obtain a mapping result. If the mapping fails, then fix the right contexts and map the left
15 contexts of the multi-lingual baseforms to obtain the mapping result. Finally, the multi-lingual context-speech mapping data is obtained according to the mapping result.

The multi-lingual anti-models are established by
20 some generation steps. First, multi-lingual query commands corresponding to certain languages are received and normalized to generate uni-lingual anti-models. The uni-lingual anti-models are then calculated to generate the multi-lingual anti-model.

25 Finally, the inventive method selects resulting speech models corresponding to the speech features from the candidate speech models according to the match probability and generates a speech command. Here, again the decision reaction engine can react to
30 the recognized speech command.

Fig. 1 is a diagram of the system for multi-lingual speech recognition according to the present invention. A system for multi-lingual speech recognition is provided. The disclosed system
5 includes a speech modeling engine 102, a speech search engine 106, and a decision reaction engine 112.

The speech modeling engine 102 receives a mixed multi-lingual speech signal 100, transfers the multi-lingual speech signal 100 into speech features 104.

10 The speech search engine 106 receives the speech features 104, and locates and compares candidate data sets 110 corresponding to the speech features 104, referring to a multi-lingual model database 108. Each of the candidate data sets 110 has several candidate
15 speech models with corresponding match probability. The locating and comparison may also refer to other rules and databases, such as a language rule database and mixed multi-lingual query commands strings. The language rule model database is established by
20 language rules of one particular field. The mixed multi-lingual query commands strings are the general terms in one particular field. The function of the mentioned reference rules and databases is to enhance the recognition rate. The speech search engine 106
25 further refers the connecting sequences of the speech models and a speech rule database 107.

The decision reaction engine 112 selects resulting speech models corresponding to the speech features from the candidate speech models according to
30 the match probability and other reference decision

rules 114. The decision reaction engine 112 then generates a speech command 116. The decision reaction engine 112 produces the reaction according to the recognized speech command 116.

5 A threshold can be designed in the reference decision rules 114 to determine whether the speech command is correctly recognized. Thus, erroneous recognized commands can be filtered, and might request reconfirm, avoiding repercussion. Otherwise, the
10 reference decision rules 114 can be designed to accept the entire recognition result without verification. The reactions may be a signal, a light, or a voice notification, prompting repeat input or an action for remote control.

15 The multi-lingual model database 108 comprises multi-lingual context-speech mapping data and multi-lingual anti-models.

Fig. 2 is a diagram of establishment of the multi-lingual context-speech mapping data according to
20 the present invention. The present invention further comprises a multi-lingual baseform mapping engine 202 and a cross-lingual diphone model generation engine 206.

The multi-lingual baseform mapping engine 202
25 compares multi-lingual query commands 200 to obtain multi-lingual baseforms. The cross-lingual diphone model generation engine 206 selects and combines the multi-lingual baseforms into the multi-lingual context-speech mapping data 208.

The cross-lingual diphone model generation engine
206 accomplishes the selection and combination by
several steps. First, the left contexts of the multi-
lingual baseforms are fixed, and the right contexts of
5 the multi-lingual baseforms are mapped to obtain a
mapping result. Next, fix the right context and the
left contexts are mapped to obtain the mapping result
if right context mapping fails. Finally, multi-
lingual context-speech mapping data is obtained
10 according to the mapping result.

Fig. 5 is a diagram illustrating cross-lingual
data of the present invention according to one
embodiment. The "c" illustrated in Fig. 5 represents
"Chinese" and the "e" represents "English." As shown
15 in the first row of Fig. 5, the "z" in Chinese cannot
generate the optimal simulated pronunciation. By
applying the provided method, the "z" in Chinese can
find diphone models "ch" or "th" for simulation, as
shown in Fig. 5. In addition, the "zcl" in Chinese
20 cannot generate the optimal mapping simulation, thus
there is no mapping result, as shown in the second row
in Fig. 5. The "ing" in Chinese maps to the
concatenation of "ih" and "ng" in English, as shown in
the third row of Fig. 5.

25 Fig. 3 is a diagram of establishment of the
multi-lingual anti-models according to the present
invention. The inventive system may include a device
32 generating the multi-lingual anti-models. The
device 32 receives multi-lingual query commands 30 to

normalize and generate all needed uni-lingual anti-model 36.

Fig. 4 is a detailed diagram of establishment of the multi-lingual anti-models according to the present invention. The device 32 comprises several uni-lingual anti-model generation engines 320, 324, 328 and an anti-model combination engine 332. The uni-lingual anti-model generation engines 320, 324, 328 receive multi-lingual query commands 30 in Fig. 3. The multi-lingual query commands 30 correspond to specific languages. The uni-lingual anti-model generation engines 320, 324, 328 normalizes specific uni-lingual diphone model database 322, 326, 330, to generate their uni-lingual anti-models. Each uni-lingual anti-model corresponds to one language. The anti-model combination engine 332, coupled to the uni-lingual anti-model generation engines 320, 324, 328, calculates the uni-lingual anti-models to generate the multi-lingual anti-models 36 in Fig. 3.

For example, the uni-lingual anti-model generation engine (language A) 320 may refer the uni-lingual diphone model database (language A) 322 to generate an uni-lingual anti-model of language A. The uni-lingual anti-model generation engine (language B) 324 may refer the uni-lingual diphone model database (language B) 326 to generate uni-lingual anti-model of language B. Similarly, the uni-lingual anti-model generation engine (language C) 328 may refer the uni-lingual diphone model database (language C) 330 to generate uni-lingual anti-model of language C. The

anti-model combination engine 332 then receives the uni-lingual anti-models of languages A, B and C and calculates them into the multi-lingual anti-model 36.

The uni-lingual anti-model generation engines
5 320, 324, 328 adopt the following formulas (1) and (2) for normalization.

$$p = \sum_{k=1} C_k N_{(o, \mu_k, \sigma_k)} \quad (1)$$

$$\log P = \text{Log} P - \log P_{anti} \quad (2)$$

If the applied speech recognition system is
10 completely trained by a mixed multi-lingual database, the system includes the trained multi-lingual dipphone models. Thus, the cross-lingual dipphone model generation engine 206 shown in Fig. 2 and the normalization shown in Fig. 4 are not necessary. If
15 the applied system is integrated by multiple independent speech recognition sub-systems, the normalization shown in Fig. 4 is required.

Fig. 6 is a diagram of an applied example of the present invention according to one embodiment. Users
20 connect to one speech recognition system through network module 616 or connecting port module 618, by which they can define desired recognition rules, such as address, ID number, or license plate number. A speech signal input 600 can be input via microphone
25 602 or telephone interface 604.

Next, the analog/digital transfer module 606 converts the speech signal input 600 into a digital signal. The programs are stored in ROM 608, and downloaded to RAM 610 and flash memory 612 for

execution at run time. The digital signal processor (DSP) unit 614 processes, controls, and recognizes data. Some fixed data, such as network protocol or boot programs, can be stored in ROM 608. The varied
5 data, such as transfer tables or speech probability models, can be stored in flash memory 612. The DSP unit 614 loads the speech recognition system into RAM 610 for data recognition.

Finally, the recognition result is sent to the
10 digital/analog module 622 for converting into analog signals. The converted analog signals could be output as an audio signal or by telephone interface 626. Moreover, the corresponding reaction 620 for the remote object, such as program upgrade or update, can
15 be executed through network module 616 or connecting port module 618.

Furthermore, the invention discloses a recognition method for multi-lingual speech recognition. Fig. 7 is a flowchart of the method for
20 multi-lingual speech recognition according to the present invention.

First, the method transfers a mixed multi-lingual speech signal into speech features.

Next, the method locates and compares candidate
25 data sets referring to a multi-lingual model database. Each of the candidate data sets corresponds to the speech features and has candidate speech models with corresponding match probability (step S702). The multi-lingual model database comprises multi-lingual

context-speech mapping data and multi-lingual anti-models.

The multi-lingual context-speech mapping data is established by a multi-lingual modeling procedure.

5 The multi-lingual modeling procedure first compares multi-lingual query commands to obtain multi-lingual baseforms. The multi-lingual modeling procedure then selects and combines the multi-lingual baseforms into multi-lingual context-speech mapping data. Selection
10 and combination are accomplished by certain steps. First, the left contexts of the multi-lingual baseforms are fixed, and the right contexts of the multi-lingual baseforms are mapped to obtain a mapping result. Next, fix the right context and the left
15 contexts are mapped to obtain the mapping result if the right context mapping fails. Finally, the multi-lingual context-speech mapping data is obtained according to the mapping result.

The multi-lingual anti-models are established by
20 a multi-lingual anti-model generation procedure. The multi-lingual anti-model generation procedure first receives multi-lingual query commands to normalizes and generate all uni-lingual anti-models. The multi-lingual anti-model generation procedure then combines
25 the uni-lingual anti-models to generate the multi-lingual anti-model.

Finally, the method selects resulting speech models corresponding to the speech features from the candidate speech models according to the match
30 probability (step S704), and generates a speech

command (step S706). The corresponding reaction may be produced according to the recognized speech command (step S708).

Thus, the system and method provided by the present invention can implement multi-lingual recognition functions to recognize multi-lingual speech signals and produce speech commands, achieving the objects of the invention. Particularly, the present invention can be applied in a speech recognition system with a large amount of vocabulary and cross-language terms, providing significant improvement over the conventional method.

It will be appreciated from the foregoing description that the system and method described herein provide a dynamic and robust solution to mixed multi-lingual speech recognition problems. If, for example, the desired language input to the system changes, the system and method of the present invention can be revised accordingly.

While the invention has been described by way of example and in terms of the preferred embodiments, it is to be understood that the invention is not limited to the disclosed embodiments. To the contrary, it is intended to cover various modifications and similar arrangements (as would be apparent to those skilled in the art). Therefore, the scope of the appended claims should be accorded the broadest interpretation so as to encompass all such modifications and similar arrangements.